Proceedings of the

First International Workshop on Free/Open-Source Rule-Based Machine Translation

2–3 November 2009
Universitat d'Alacant
Alacant, Spain

Edited by
Juan Antonio Pérez-Ortiz
Felipe Sánchez-Martínez
Francis M. Tyers

Preface

It is our pleasure to welcome you to the First International Workshop on Free/Open-Source Rule-Based Machine Translation to be held in Alacant, Spain. This workshop has been inspired, on one hand, by the successful OSMaTran workshop on open-source machine translation —held on September, 2005 at the Tenth Machine Translation Summit— where 3 of the 4 presented papers where about rule-based machine translation; and, on the other, by the successful participation of the Apertium project in the Google Summer of Code 2009.

The free/open-source software movement has arrived into the field of machine translation. Machine translation is special in that, in addition to specific algorithms, it heavily depends on extensive language-dependent data. Therefore, not only the engine or the tools used to manage these data have to be free/open-source, but also the data themselves. There are many machine translation packages of this type available; however most of them are corpus-based, and, in particular, statistical machine translation systems. Rule-based machine translation systems built on these principles are still little known and little used.

There are distinct advantages of having free/open-source licences for rule-based machine translation: the linguistic knowledge for the translation of one language into another, which is explicitly encoded in the form of linguistic data so that both humans and the machine translation engine can process it, can be reused to build data for other language pairs or even for other human language technologies besides machine translation, and, conversely, linguistic knowledge from other sources may be reused to build machine translation systems. The free and open scenario makes this reuse easier, and, if *copyleft* licences are used, builds a commons of knowledge and resources that benefits all the language communities involved. These advantages are even clearer for less-resourced languages, for which large bilingual corpora are not available, and for morphologically rich languages, which even with large corpora suffer from data sparseness.

This workshop aims at bringing together the experience of researchers and developers in the field of rule-based machine translation who have decided to board the free/open-source train and are effectively contributing to create that commons of explicit knowledge: machine translation rules and dictionaries, and machine translation systems whose behaviour is transparent and clearly traceable through their explicit logic. Each of the 16 papers submitted from 11 countries was peer-reviewed by two independent reviewers from the program committee, resulting in a selection of the 10 papers included in these proceedings.

Our special thanks goes to Dr. Amba Kulkarni, from the University of Hyderabad, and Dr. Kepa Sarasola, from Euskal Herriko Unibertsitatea, who kindly agreed to give keynote addresses at the workshop. We also acknowledge all the reviewers, whose names are subsequently listed, for their detailed extensive reviews and useful recommendations which where vital in helping authors to improve their papers. We also appreciate the dedicated efforts of the local organising committee who worked hard to plan this workshop. We are also very grateful to the sponsors who supported this workshop: Institut Universitari d'Investigació Informàtica

of Universitat d'Alacant. Departament de Llenguatges i Sistemes Informàtics of Universitat d'Alacant, Prompsit Language Engineering, and also to the people responsible for the Google Summer of Code 2009, who let us invest the grant received by the Apertium project in planning this workshop. Finally, a big thank you goes to all the authors who made this workshop both possible and successful.

All the papers included in this proceedings can be found at the Open Access Repository of Universitat d'Alacant on `http://rua.ua.es/dspace/handle/10045/11809`.

We look forward to an exciting and productive workshop.

Alacant, November 2009

Juan Antonio Pérez-Ortiz and **Felipe Sánchez-Martínez**
FreeRBMT09 Programme Committee co-chairs

Editors:

Juan Antonio Pérez-Ortiz, Universitat d'Alacant
Francis M. Tyers, Universitat d'Alacant
Felipe Sánchez-Martínez, Universitat d'Alacant

Programme Committee Chairs:

Juan Antonio Pérez-Ortiz, Universitat d'Alacant
Felipe Sánchez-Martínez, Universitat d'Alacant

Programme Committee:

Mikel L. Forcada, Universitat d'Alacant and Dublin City University
Hrafn Loftsson, Háskólinn í Reykjavík
Jacob Nordfalk, Ingeniørhøjskolen i København
Lluís Padró, Universitat Politècnica de Catalunya
Kepa Sarasola, Euskal Herriko Unibertsitatea
Kevin P. Scannell, Saint Louis University
Trond Trosterud, Romssa Universitehta

Additional Reviewers:

Aingeru Mayor, Euskal Herriko Unibertsitatea
Sudip Naskar, Dublin City University

Invited Speakers:

Amba Kulkarni, University of Hyderabad
Kepa Sarasola, Euskal Herriko Unibertsitatea

Local Organising Committee:

Miquel Esplà-Gomis, Universitat d'Alacant
Xavier Ivars-Ribes, Universitat d'Alacant
Juan Antonio Pérez-Ortiz, Universitat d'Alacant
Víctor M. Sánchez-Cartagena, Universitat d'Alacant
Felipe Sánchez-Martínez, Universitat d'Alacant
Francis M. Tyers, Universitat d'Alacant

Sponsoring Institutions:

Institut Universitari d'Investigació Informàtica, Universitat d'Alacant
Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant
Prompsit Language Engineering, S.L.
Transducens Research Group, Universitat d'Alacant

Table of Contents

Invited talks

Matxin: developing sustainable machine translation for a less-resourced language
Kepa Sarasola 1

Anusaaraka: An accessor cum machine translator
Amba Kulkarni 1

Platforms for free/open-source rule-based machine translation

The Apertium machine translation platform: Five years on
Mikel L. Forcada, Francis M. Tyers and Gema Ramírez-Sánchez 3

Matxin: Moving towards language independence
Aingeru Mayor and Francis M. Tyers 11

OpenLogos MT and the SAL representation language
Bernard Scott and Anabela Barreiro 19

Free/open-source linguistic data and language pairs

Shallow-transfer rule-based machine translation for Swedish to Danish
Francis M. Tyers and Jacob Nordfalk 27

Reuse of free resources in machine translation between Nynorsk and Bokmål
Kevin Unhammer and Trond Trosterud 35

Development of a morphological analyser for Bengali
Abu Zaher Md. Faridee and Francis M. Tyers 43

Technical issues in free/open-source machine translation

An open-source highly scalable web service architecture for the Apertium machine translation engine
Víctor M. Sánchez-Cartagena and Juan Antonio Pérez-Ortiz 51

Apertium goes SOA: an efficient and scalable service based on the Apertium rule-based machine translation platform
Pasquale Minervini 59

A trigram part-of-speech tagger for the Apertium free/open-source machine translation Platform
Zaid Md Abdul Wahab Sheikh and Felipe Sánchez-Martínez 67

Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya
Luis Villarejo Muñoz, Sergio Ortiz Rojas and Mireia Ginestí Rosell 75

Workshop Programme

Monday, 2nd November 2009

09:00 **Reception**

09:15 **Opening**

09:30 **Invited talk**: *Matxin: developing sustainable machine translation for a less-resourced language*
Kepa Sarasola, Euskal Herriko Unibertsitatea, Spain

10:30 **Coffee**

Session: **Platforms for free/open-source rule-based machine translation**

11:00 *The Apertium machine translation platform: Five years on*
Mikel L. Forcada, Francis M. Tyers and Gema Ramírez-Sánchez

11:45 *Matxin: Moving towards language independence*
Aingeru Mayor and Francis M. Tyers

12:30 *OpenLogos MT and the SAL representation language*
Bernard Scott and Anabela Barreiro

13:15 **Lunch**

Session: **Free/open-source linguistic data and language pairs**

15:15 *Shallow-transfer rule-based machine translation for Swedish to Danish*
Francis M. Tyers and Jacob Nordfalk

15:45 *Reuse of free resources in machine translation between Nynorsk and Bokmål*
Kevin Unhammer and Trond Trosterud

16:15 *Development of a morphological analyser for Bengali*
Abu Zaher Md. Faridee and Francis M. Tyers

Tuesday, 3rd November 2009

09:30	**Invited talk**:	*Anusaaraka: An accessor cum machine translator* Amba Kulkarni, University of Hyderabad, India
10:30	**Coffee**	
	Session:	**Technical issues in free/open-source machine translation**
11:00		*An open-source highly scalable web service architecture for the Apertium machine translation engine* Víctor M. Sánchez-Cartagena and Juan Antonio Pérez-Ortiz
11:30		*Apertium goes SOA: an efficient and scalable service based on the Apertium rule-based machine translation platform* Pasquale Minervini
12:00		*A trigram part-of-speech tagger for the Apertium free/open-source machine translation platform* Zaid Md. Abdul Wahab Sheikh and Felipe Sánchez-Martínez
12:30		*Joint efforts to further develop and incorporate Apertium into the document management flow at the Universitat Oberrta de Catalunya* Luis Villarejo Muñoz, Sergio Ortiz Rojas and Mireia Ginestí Rosell
13:00	**Lunch**	
15:00	**Demos**	
16:00	**Closing**	
16:15	**Meeting:**	Extraordinary meeting of the Apertium machine translation project

Invited talks

Matxin: developing sustainable machine translation for a less-resourced language

Kepa Sarasola, Euskal Herriko Unibertsitatea, Spain

Following the strategy defined in IXA group for reusing linguistic resources and NLP tools, in year 2000 (but not before), we decided that we had enough languages resources and tools (bilingual dictionaries, morphological and syntactic analysers and parsers) that could be reused to build an RBMT system for the Spanish–Basque pair. The system built is called Matxin and is available at `matxin.sourceforge.net`. Since 2006 we are collaborating with DCU building a Spanish–Basque system based on EBMT and SMT paradigms. We could get better results with bigger parallel corpus, but it is difficult to get it for Basque, a minority language. Based on our work we have published a strategy for sustainable MT for lesser-resourced languages; it is based on incremental design, reusability, standardisation and open source. We have developed MT engines based on the three paradigms (RBMT, SMT and EBMT), so our position is optimal to experiment with hybrid systems and multi-engine systems.

Anusaaraka: An accessor cum machine translator

Amba Kulkarni, University of Hyderabad, India

India being multilingual, there is a demand for translation both among Indian languages as well as from English to Indian languages. Translation being not reliable, Anusaaraka aims to provide complete access to the source text in addition to translation. With an appropriate division of load between man and machine, Kannada-Hindi Anusaaraka, developed in early 90s, demonstrated that it is possible to reduce the language barrier considerably. However it is necessary for an Anusaaraka reader to undergo some training on the syntactic divergences and special notation used to handle the divergences in word-meaning mappings between the source and the target language. In the later version of Anusaaraka, in order to reduce the burden on a user, the state-of-the-art MT system formed an important component of it. Care was taken to develop the architecture in such a way that, it can cater to the needs of diverse requirements ranging from faithful access to the full fledged translation.